DOI:
10.1214/18-AIHP946

Document Version
Peer reviewed version

Asymptotic nonequivalence of density estimation and Gaussian white noise for small densities

Kolyan Ray* and Johannes Schmidt-Hieber
King's College London and Leiden University

Abstract

It is well-known that density estimation on the unit interval is asymptotically equivalent to a Gaussian white noise experiment, provided the densities are sufficiently smooth and uniformly bounded away from zero. We show that a uniform lower bound, whose size we sharply characterize, is in general necessary for asymptotic equivalence to hold.

AMS 2010 Subject Classification: Primary 62B15; secondary 62G07, 62G10, 62G20.

Keywords: Asymptotic equivalence; density estimation; Gaussian white noise model; small densities.

1 Introduction

A fundamental problem in nonparametric statistics is density estimation on a compact set, say the unit interval $[0,1]$, where we observe n i.i.d. observations from an unknown probability density f. If the parameter space Θ consists of densities f that are uniformly bounded away from zero and have Hölder smoothness $\beta > 1/2$, then a seminal result of Nussbaum [18] establishes the global asymptotic equivalence of this experiment to the Gaussian white noise model where we observe $(Y_t)_{t\in[0,1]}$ arising from

$$dY_t = 2\sqrt{f(t)}dt + n^{-1/2}dW_t, \quad t \in [0,1], \quad f \in \Theta, \tag{1}$$

*The research leading to these results has received funding from the European Research Council under ERC Grant Agreement 320637.

Email: kolyan.ray@kcl.ac.uk, schmidthieberaj@math.leidenuniv.nl

where $(W_t)_{t\in[0,1]}$ is a Brownian motion. The smoothness constraint is sharp: Brown and Zhang [4] construct a counterexample with a parameter space of Hölder smoothness exactly $\beta = 1/2$ such that asymptotic equivalence does not hold.

If two statistical experiments are asymptotically equivalent in the Le Cam sense, then asymptotic statements can be transferred between the experiments. More precisely, the existence of a decision procedure with risk R_n for a given bounded loss function in one model implies the existence of a corresponding decision procedure with risk $R_n + o(1)$ in this loss in the other model. To derive asymptotic properties, one may therefore work in the simpler model and transfer the results to the more complex model. This is one of the main motivations behind the study of asymptotic equivalence. The last part of the introduction provides definitions and summarizes the concept of asymptotic equivalence of statistical experiments.

In practice, densities may be small or even zero on a subset of the domain, in which case the above result no longer applies. The goal of this article is to contribute to the general understanding of necessary conditions for asymptotic equivalence to hold, in particular the necessity of uniform boundedness away from zero. We show that without a minimal lower bound on the densities, density estimation and the Gaussian model (1) are always asymptotically nonequivalent, irrespective of the amount of Hölder smoothness.

In fact, we prove a more precise result by characterizing a size threshold such that if densities fall below this level, asymptotic equivalence never holds. In a companion paper [22], we show constructively that above this threshold, asymptotic equivalence may still hold. Our threshold is thus sharp in the sense that it is the smallest possible value a density can take such that asymptotic equivalence can hold.

We employ sample-size dependent parameter spaces $\Theta = \Theta_n$, as is typical in high-dimensional statistics. We prove that if the parameter spaces contain a sequence of β-smooth densities $(f_n)_n$ such that $\inf_{x\in[0,1]} f_n(x) = O(n^{-\beta/(\beta+1)})$ for all n, as well as suitable neighbourhoods of β-smooth densities around the (f_n), then the experiments are always asymptotically nonequivalent. This is a natural threshold for describing "small" and "large" densities with, for instance, different minimax rates attainable above and below this level [19, 20], see (2) and the related discussion below.

From a practical perspective, Gaussian approximations have been proposed in density estimation (e.g. [1]) and one would like to better understand how "large" a density must be for such methods to be applicable. The present work is a step in this direction. Furthermore, all the results presented in this paper also hold for the closely related case of Poisson intensity estimation, which is always asymptotically equivalent to density estimation, irrespective of

density size or Hölder smoothness [15, 22]. This case is of particular practical relevance given the widespread use of Gaussian approximations for Poisson data [12], even for small intensities [16]. We avoid further mention of Poisson intensity estimation for conciseness, but readers should bear in mind that all the present results and conclusions apply equally to that model.

There are few results establishing the necessity of conditions for asymptotic equivalence via counterexamples. For nonparametric regression, [4, 8] show the necessity of smoothness assumptions. The paper [26] establishes nonequivalence between the GARCH model and its diffusion limit under stochastic volatility, as well as their equivalence under deterministic volatility.

The proof we employ here relies on a reduction to binary experiments. The difficulty lies in both the construction of a suitable two-point testing problem and also in obtaining sufficiently good bounds on the total variation distance. Indeed, the situation is rather more subtle than one might first imagine. For two-point hypothesis testing problems, we show that one can consistently test between the alternatives in one model if and only if one can do so in the other (Lemma 2). To establish nonequivalence, one must therefore construct alternatives which can be separated with a positive probability that is strictly bounded away from zero and one and for which suitable bounds can be computed.

Although for small signals, density estimation and the Gaussian white noise model (1) are no longer asymptotically equivalent, many aspects of their statistical theory remain the same. As mentioned above, simple hypothesis testing is essentially the same in both models without any lower bound on the densities. To explain this in more detail, suppose that g_n and h_n are two sequences of densities and denote the probability measures in the density estimation model and the Gaussian model (1) by P_f^n and Q_f^n respectively. The sums of the type I and II error probabilities of the Neyman-Pearson test for the simple hypotheses

$$H_0 : f = g_n \qquad H_1 : f = h_n$$

in the two models are $\frac{1}{2}(1 - \|P_{g_n}^n - P_{h_n}^n\|_{\mathrm{TV}})$ and $\frac{1}{2}(1 - \|Q_{g_n}^n - Q_{h_n}^n\|_{\mathrm{TV}})$ respectively. By Lemma 2 below, $\|P_{g_n}^n - P_{h_n}^n\|_{\mathrm{TV}} \to 1$ if and only if $\|Q_{g_n}^n - Q_{h_n}^n\|_{\mathrm{TV}} \to 1$, which shows that we can consistently test against a simple alternative in one model if and only if we can do so in the other model. This argument requires no lower bound on the densities. The Hellinger distance also behaves very similarly in the two models, see Lemma 2 for a precise statement. It is an interesting phenomenon that while the models are potentially far apart with respect to the Le Cam distance, information distances, such as the total variation and Hellinger distance, remain close. Although this does not hold for all common information measures, for instance the Kullback-Leibler divergence, it nevertheless suggests that negative results

for small densities in the Le Cam sense may be misleading, since many important statistical properties still carry over between models.

Beyond density estimation, uniform boundedness away from zero is a standard assumption in the asymptotic equivalence literature [2, 9, 10, 11, 18]. However, this assumption is not always required, including in regression type models [3, 23] and even some non-linear problems, such as diffusion processes [5, 6, 7]. A better understanding of the necessity of such conditions is therefore of interest in a wide variety of models.

2 Main results

Basic notation and definitions

For two functions f, g on $[0,1]$, we write $f \leq g$ if $f(x) \leq g(x)$ for all $x \in [0,1]$ and let $\|f\|_2$ denote the L^2-norm of f. Given two probability measures P, Q with densities p, q with respect to some dominating measure ν, we recall the total variation distance $\|P - Q\|_{\mathrm{TV}} := \frac{1}{2}\int |p - q| d\nu$ and Hellinger distance $H(P,Q) := (\int (\sqrt{p} - \sqrt{q})^2 d\nu)^{1/2}$.

A statistical experiment $\mathcal{E}(\Theta) = (\Omega, \mathcal{A}, (P_\theta : \theta \in \Theta))$ consists of a sample space Ω with associated σ-algebra $\mathcal{A}$ and a family $(P_\theta : \theta \in \Theta)$ of probability measures all defined on the measurable space $(\Omega, \mathcal{A})$. We call $\mathcal{E}(\Theta)$ dominated if there exists a probability measure μ such that any P_θ is dominated by μ. Furthermore, $\mathcal{E}(\Theta)$ is said to be Polish if Ω is a Polish space and $\mathcal{A}$ is the associated Borel σ-algebra. If $\mathcal{E}(\Theta) = (\Omega, \mathcal{A}, (P_\theta : \theta \in \Theta))$ and $\mathcal{F}(\Theta) = (\Omega', \mathcal{A}', (Q_\theta : \theta \in \Theta))$ are two Polish and dominated experiments indexed by the same parameter space, the Le Cam deficiency can be defined as

$$\delta\big(\mathcal{E}(\Theta), \mathcal{F}(\Theta)\big) := \inf_M \sup_{\theta \in \Theta} \big\| M P_\theta^n - Q_\theta^n \big\|_{\mathrm{TV}},$$

where the infimum is taken over all Markov kernels $M : \Omega_1 \times \mathcal{A}_2 \to [0,1]$. The Le Cam distance is defined as

$$\Delta\big(\mathcal{E}(\Theta), \mathcal{F}(\Theta)\big) := \max\big\{\delta\big(\mathcal{E}(\Theta), \mathcal{F}(\Theta)\big), \delta\big(\mathcal{F}(\Theta), \mathcal{E}(\Theta)\big)\big\},$$

which defines a pseudo-distance on the space of all experiments with parameter space Θ. One may generalize the definition of Le Cam deficiency to spaces that are neither Polish nor dominated upon replacing the notion of Markov kernel with a more general transition [14, 24]. However, we refrain from doing so here since these notions coincide in the Polish and dominated experiments we consider in this article, see (68) and Proposition 9.2 of [18]. Finally, we say that two sequences of experiments $\mathcal{E}_n(\Theta_n) = (\Omega_n, \mathcal{A}_n, (P_\theta^n : \theta \in \Theta_n))$ and

$\mathcal{F}(\Theta_n) = (\Omega'_n, \mathcal{A}'_n, (Q^n_\theta : \theta \in \Theta_n))$ are asymptotically equivalent if $\Delta(\mathcal{E}_n(\Theta_n), \mathcal{F}_n(\Theta_n)) \to 0$ as $n \to \infty$. General treatments on asymptotic equivalence can be found in [14, 24].

In this article, we consider the following two statistical experiments.

Density estimation $\mathcal{E}^D_n(\Theta)$*:* In nonparametric density estimation, we observe n i.i.d. copies $X_1, \ldots, X_n$ of a random variable on $[0,1]$ with unknown Lebesgue density f. The corresponding statistical experiment is $\mathcal{E}^D_n(\Theta) = ([0,1]^n, \sigma([0,1]^n), (P^n_f : f \in \Theta))$ with P^n_f the product probability measure of $X_1, \ldots, X_n$.

Gaussian white noise experiment $\mathcal{E}^G_n(\Theta)$*:* We observe the Gaussian process $(Y_t)_{t\in[0,1]}$ arising from (1) with $f \in \Theta$ unknown. Denote by $\mathcal{C}([0,1])$ the space of continuous functions on $[0,1]$ and let $\sigma(\mathcal{C}([0,1]))$ be the σ-algebra generated by the open sets with respect to the uniform norm. The Gaussian white noise experiment is then given by $\mathcal{E}^G_n(\Theta) = (\mathcal{C}([0,1]), \sigma(\mathcal{C}([0,1])), (Q^n_f : f \in \Theta))$ with Q^n_f the distribution of $(Y_t)_{t\in[0,1]}$.

Function spaces

Denote by $\lfloor\beta\rfloor$ the largest integer strictly smaller than β. The usual Hölder semi-norm is given by $|f|_{\mathcal{C}^\beta} := \sup_{x\neq y, x,y\in[0,1]} |f^{(\lfloor\beta\rfloor)}(x) - f^{(\lfloor\beta\rfloor)}(y)|/|x-y|^{\beta-\lfloor\beta\rfloor}$ and the Hölder norm is $\|f\|_{\mathcal{C}^\beta} := \|f\|_\infty + \|f^{(\lfloor\beta\rfloor)}\|_\infty + |f|_{\mathcal{C}^\beta}$. Consider the space of β-smooth Hölder densities with Hölder norm bounded by R,

$$\mathcal{C}^\beta(R) := \big\{ f : [0,1] \to \mathbb{R} \ : \ f \geq 0, \ \int_0^1 f(u)du = 1, \ f^{(\lfloor\beta\rfloor)} \text{ exists}, \ \|f\|_{\mathcal{C}^\beta} \leq R \big\}.$$

If $0 < \beta \leq 2$, the pointwise rate of estimation at any $x \in (0,1)$ over the parameter space $\mathcal{C}^\beta(R)$ is given by

$$n^{-\frac{\beta}{\beta+1}} + \left(\frac{f(x)}{n}\right)^{\frac{\beta}{2\beta+1}}, \tag{2}$$

with upper and lower bounds matching up to $\log n$ factors (see Theorems 3.1 and 3.3 of [19] for density estimation and Theorems 1 and 2 of [20] for the Gaussian white noise model). There is thus a phase transition in the estimation rate for small densities occurring at the n-dependent signal size $f(x) \asymp n^{-\frac{\beta}{\beta+1}}$. This is the same boundary for asymptotic nonequivalence proved in Theorem 1 below, so that in some respects at least, the two experiments do behave differently from one another below this threshold. However, despite asymptotic nonequivalence, many other properties, such as minimax rates and consistent testing, are still asymptotically the same below this threshold. Indeed, the counterexample we construct lies right on the boundary of testing problems and in some sense only narrowly fails. The importance of the threshold $f(x) \asymp n^{-\frac{\beta}{\beta+1}}$ is not isolated to minimax estimation rates and asymptotic equivalence and seems to play a fundamental role for small densities,

for example being necessary to obtain sharp rates when estimating the support of a density [19]. For further discussion see [19, 20].

The rate of convergence (2) does not extend to $\beta > 2$ using the usual definition of Hölder smoothness due to the existence of functions which are highly oscillatory near zero (Theorem 3 of [20]). A natural way to attain the rate for smoothness $\beta > 2$ is to impose a shape constraint ruling out such pathological behaviour. On $\mathcal{C}^\beta$, define the flatness seminorm

$$|f|_{\mathcal{H}^\beta} = \max_{1 \le j < \beta} \| |f^{(j)}|^\beta / |f|^{\beta - j} \|_\infty^{1/j} = \max_{1 \le j < \beta} \left(\sup_{x \in [0,1]} \frac{|f^{(j)}(x)|^\beta}{|f(x)|^{\beta - j}} \right)^{1/j} \tag{3}$$

with $0/0$ defined as 0 and $|f|_{\mathcal{H}^\beta} = 0$ if $0 < \beta \le 1$. The quantity $|f|_{\mathcal{H}^\beta}$ measures the flatness of a function near zero in the sense that if $f(x)$ is small, then the derivatives of f must also be small in a neighborhood of x. Define $\|f\|_{\mathcal{H}^\beta} := \|f\|_{\mathcal{C}^\beta} + |f|_{\mathcal{H}^\beta}$ and consider the space of densities

$$\mathcal{H}^\beta(R) := \{ f \in \mathcal{C}^\beta(R) \ : \ \|f\|_{\mathcal{H}^\beta} \le R \}.$$

Notice that $\mathcal{H}^\beta(R) = \mathcal{C}^\beta(R)$ for $\beta \le 1$. For further discussion and properties of the function space $\mathcal{H}^\beta(R)$, see [21].

The reason we construct a counterexample in $\mathcal{H}^\beta(R)$ is to concretely show that asymptotic nonequivalence is not due to functions that are highly oscillatory near zero, but also holds for typical Hölder functions. Thus even when considering only "nice" Hölder functions, for which the rate (2) is attainable, nonequivalence still holds.

Asymptotic nonequivalence

To obtain suitable lower bounds on the Le Cam deficiencies, we require that the small densities are not isolated in the parameter space Θ_n, meaning we must introduce a notion of interior parameter space. This is in some sense necessary, since asymptotic equivalence may still hold when the small density behaviour is driven by a parametric component, in particular having finite Hellinger metric dimension. For further discussion on this point, see Proposition 1 below.

The following result is the main contribution of this article, showing that if

$$\inf_{f \in \Theta_n} \inf_{x \in [0,1]} f(x) \lesssim n^{-\beta/(\beta+1)},$$

then the Le Cam deficiency is bounded from below by a positive constant for sufficiently large n. In this case, the experiments are asymptotically nonequivalent.

Theorem 1. *Let $R, \beta > 0$. There exists a constant $c > 1$, not depending on R, such that if $(f_{0,n})_n \subset \Theta_n \cap \mathcal{H}^\beta(R)$ is a sequence satisfying $\inf_{x\in[0,1]} f_{0,n}(x) \le \frac{1}{2}R^{1/(\beta+1)}n^{-\beta/(\beta+1)}$ and $\{f \in \mathcal{H}^\beta(cR) : c^{-1}f_{0,n} \le f \le cf_{0,n}\} \subset \Theta_n$ for all $n \ge 2$, then*

$$\delta\big(\mathcal{E}_n^D(\Theta_n), \mathcal{E}_n^G(\Theta_n)\big) \ge 0.007 + o(1) > 0.$$

The assumption is that the parameter space Θ_n is rich enough to contain a function $f_{0,n}$ that somewhere falls below the threshold $n^{-\beta/(\beta+1)}$, together with all the functions in $\mathcal{H}^\beta(cR)$ lying in the band $x \mapsto [c^{-1}f_{0,n}(x), cf_{0,n}(x)]$ around $f_{0,n}$. An explicit expression for c can be obtained from the proof, see (15). As a particular example, the norm balls $\Theta_n = \mathcal{H}^\beta(R)$ satisfy the above assumptions.

Corollary 1. *For any $\beta > 0$ and sufficiently large $R \ge R_0(\beta)$,*

$$\Delta\big(\mathcal{E}_n^D(\mathcal{H}^\beta(R)), \mathcal{E}_n^G(\mathcal{H}^\beta(R))\big) \ge 0.007 + o(1) > 0.$$

Proof of Corollary 1. For $\beta > 0$, consider the density $f(x) = (\beta+1)x^\beta$. For any integer $0 \le j < \beta$, $f^{(j)}(x) = [\Gamma(\beta+2)/\Gamma(\beta-j+1)]x^{\beta-j}$, where $\Gamma(t)$ denotes the Gamma function. This implies that $\|f\|_\infty + \|f^{(\lfloor\beta\rfloor)}\|_\infty \le (\beta+1) + \Gamma(\beta+2)/\Gamma(\beta-\lfloor\beta\rfloor+1)$ and $|f|_{\mathcal{H}^\beta} = \max_{1\le j<\beta}\Gamma(\beta+2)^{\beta/j}\Gamma(\beta-j+1)^{-\beta/j}(\beta+1)^{-(\beta-j)/j}$. For any $z \in [0,1]$ and $\gamma \in [0,1]$, $1 \le z^\gamma + (1-z)^\gamma$, which implies $|x^\gamma - y^\gamma| \le |x-y|^\gamma$ for $x, y \ge 0$. Hence, $|f|_{\mathcal{C}^\beta} \le \Gamma(\beta+2)/\Gamma(\beta-\lfloor\beta\rfloor+1)$, so that $\|f\|_{\mathcal{H}^\beta} \le C_\beta$ for some finite constant C_β depending only on β. Let $c > 1$ be the constant in Theorem 1. If $R \ge C_\beta c$, we may apply Theorem 1 with the constant sequence $f_{0,n} = f$, since then $(f_{0,n})_n \subset \mathcal{H}^\beta(C_\beta) \subset \mathcal{H}^\beta(R/c)$. By Theorem 1 with $\Theta_n = \mathcal{H}^\beta(R)$ and R replaced by R/c, the assertion follows. □

Since $\|f\|_{\mathcal{H}^\beta} \ge \|f\|_\infty \ge 1$ for any density f on $[0,1]$, the radius R in the previous corollary must be larger than some $R_0(\beta)$, otherwise the parameter space $\mathcal{H}^\beta(R)$ is empty. For small densities, the Gaussian white noise model (1) can be asymptotically more informative than density estimation. This result is only interesting in the case $\beta > 1/2$, since for $\beta \le 1/2$, asymptotic equivalence can fail even if all densities are uniformly bounded away from zero [4].

Under general conditions, if $\Theta_n \subset \mathcal{H}^\beta(R)$ for $1/2 < \beta \le 1$ and $\inf_{f\in\Theta_n}\inf_{x\in[0,1]} f(x) \gg n^{-\frac{\beta}{\beta+1}}\log^8 n$, the squared Le Cam deficiencies between density estimation and the Gaussian model (1) are exactly of the order

$$\min\left\{1, n^{\frac{1-2\beta}{2\beta+1}} \sup_{f\in\Theta_n}\int_0^1 f(x)^{-\frac{2\beta+3}{2\beta+1}}dx\right\}, \tag{4}$$

see Theorem 4 of [22]. In particular, if f is uniformly bounded away from zero we recover the rate $\min\{1, n^{(1-2\beta)/(2\beta+1)}\}$, so that the experiments are asymptotically equivalent if and only if $\beta > 1/2$. As we now show by example, in view of (4), the threshold $n^{-\beta/(\beta+1)}$ obtained in Theorem 1 is essentially sharp up to a logarithmic factor.

Consider the densities $f_{0,n}(x) \propto x^\beta + n^{-\frac{\beta}{\beta+1}} M_n$ with $M_n \gtrsim \log^8 n$ diverging, which satisfy $\inf_{x\in[0,1]} f_{0,n}(x) \propto n^{-\frac{\beta}{\beta+1}} M_n \gg n^{-\frac{\beta}{\beta+1}}$ and $f_{0,n} \in \mathcal{H}^\beta(R)$ for $R > 0$ large enough. For $c > 0$ the constant from Theorem 1, set

$$\Theta_n = \big\{ f \in \mathcal{H}^\beta(cR) : c^{-1} f_{0,n} \le f \le c f_{0,n} \big\}.$$

Since $\inf_{f\in\Theta_n} \inf_{x\in[0,1]} f(x) \gtrsim n^{-\frac{\beta}{\beta+1}} \log^8 n$, applying (4),

$$\Delta\big(\mathcal{E}_n^D(\Theta_n), \mathcal{E}_n^G(\Theta_n)\big)^2 \asymp n^{\frac{1-2\beta}{2\beta+1}} \int_0^1 f_{0,n}(x)^{-\frac{2\beta+3}{2\beta+1}} dx \asymp M_n^{\frac{(1-2\beta)(\beta+1)}{\beta(2\beta+1)}} \to 0$$

for $1/2 < \beta \le 1$, so that density estimation and the Gaussian model (1) with parameter spaces Θ_n are asymptotically equivalent. In summary, asymptotic equivalence always fails below the threshold $n^{-\frac{\beta}{\beta+1}}$, but may still hold for any level larger than $n^{-\frac{\beta}{\beta+1}} \log^8 n$, thereby showing that Theorem 1 is sharp up to a logarithmic factor. The $\log^8 n$ factor is a technical artifact arising from the proof of (4).

Asymptotic equivalence for small densities in parametric settings

Asymptotic nonequivalence due to small densities is a feature of fully nonparametric models and our conclusions do not necessarily apply in parametric models. We illustrate this via an example, whose proof we defer to the end of the article.

Proposition 1. *Consider the probability density* $g(x) = 960x^2(1/2 - x)^2 1_{[0,1/2]}(x)$. *For* $K \subset (0, 1/2)$ *a compact interval, consider the location family* $\Theta(K) = \{f_\theta(x) = g(x - \theta) : \theta \in K\}$. *For this parameter space, density estimation and the Gaussian model* (1) *are asymptotically equivalent, that is as* $n \to \infty$,

$$\Delta\big(\mathcal{E}_n^D(\Theta(K)), \mathcal{E}_n^G(\Theta(K))\big) \to 0.$$

The densities in the location family $\Theta(K)$ are not bounded away from zero on $[0, 1]$, with f_θ equal to zero on $[0, \theta] \cup [\theta + 1/2, 1]$, yet asymptotic equivalence still holds. The reason for this is that for areas of $[0, 1]$ where there are too few observations to admit a Gaussian approximation, the required information is provided by the parameter estimates for θ. A sufficient condition for this is finite Hellinger *metric dimension* in the density model, not to be confused with finite vectorial dimension of the parameter space, see Assumption (A3)

of Le Cam [13]. Recall that a family $(f_\theta : \theta \in \Theta')$ of density functions is said to have finite Hellinger metric dimension if there exists a number $D \geq 0$ such that every subset of $(f_\theta : \theta \in \Theta')$ which can be covered by an ε-ball in Hellinger distance H, can be covered by at most 2^D $\varepsilon/2$-balls in H, where D does not depend on ε. For example, the family of densities $\{C\exp(-|x-\theta|^\alpha) : \theta \in \mathbb{R}\}$ on $\mathbb{R}$ for some $\alpha \in (0, 1/2)$ has vectorial dimension one yet does not have finite Hellinger metric dimension, see Remark 2 after Theorem 4.3 of Le Cam [13]. In this sense, Theorem 1 is truly a nonparametric result.

One can extend this further by considering parameter spaces with a-priori known zeroes. For instance Mariucci [17] establishes asymptotic equivalence for densities of the form fg, where $f \in C^\beta$, $\beta > 1$, is an unknown function uniformly bounded away from zero and g is a given known function that is possibly small. In view of the above, one may interpret this as a form of semiparametric model, with a parametric part g determining the density for small values and the nonparametric part f doing so for large values. Thus for areas of $[0, 1]$ with sufficient observations, one can fit a Gaussian approximation based on f as usual, whereas for regions with insufficient observations, one must use the information provided by the parameter estimate for g, which in this particular example arises from a zero-dimensional family since g is known exactly.

Overview of the proof

The proof of Theorem 1 is based on a reduction to binary experiments and a direct comparison of the total variation distances between the parameters using the following lemma.

Lemma 1. *Let* $\mathcal{E}_1^b = (\Omega_1, \mathcal{A}_1, (P_{1,i} : i \in \{1, 2\}))$ *and* $\mathcal{E}_2^b = (\Omega_2, \mathcal{A}_2, (P_{2,i} : i \in \{1, 2\}))$ *be binary experiments. Then*

$$\delta(\mathcal{E}_1^b, \mathcal{E}_2^b) \geq \frac{1}{2}\big(\|P_{2,1} - P_{2,2}\|_{\mathrm{TV}} - \|P_{1,1} - P_{1,2}\|_{\mathrm{TV}}\big).$$

Proof. We have the explicit formula $\delta(\mathcal{E}_1^b, \mathcal{E}_2^b) = \sup_{0\leq\xi\leq1}[g_1(\xi) - g_2(\xi)]$ with $g_j(\xi) = \inf[(1-\xi)P_{j,1}\phi + \xi P_{j,2}(1-\phi)]$ the error function in $\mathcal{E}_j^b$, $j \in \{1, 2\}$, and where the infimum is over all tests ϕ, see Strasser [24], Corollary 15.7 and Definition 14.1. Notice that the definition of deficiency in [24], Definition 15.1, has an additional factor $1/2$. The result then follows with $g_j(1/2) = \frac{1}{2}(1 - \|P_{j,1} - P_{j,2}\|_{\mathrm{TV}})$ ([24], p. 71). □

To establish asymptotic nonequivalence for a discrete experiment and its continuous analogue, a standard approach is to consider a sequence of binary experiments such that the total variation distance in the discrete model is zero (i.e. both measures are the same)

but the total variation distance in the continuous model is positive. Lemma 1 then yields asymptotic nonequivalence.

This approach cannot be used here and the proof of Theorem 1 requires a much more careful choice of the sequence of binary experiments. Consider a sequence of binary experiments $\{P_{f_n}^n, P_{g_n}^n\}$ in the density estimation setting with corresponding binary experiments $\{Q_{f_n}^n, Q_{g_n}^n\}$ in the Gaussian white noise model. The following result shows that the total variation distance in one experiment tends to zero if and only if the total variation in the other experiment also tends to zero. The same holds if the total variation distances both tend to one. Thus, in order to construct a lower bound via Lemma 1, such sequences cannot be used.

Lemma 2. *Let $(f_n)_n$ and $(g_n)_n$ be arbitrary sequences of densities in both experiments $\mathcal{E}_n^D(\Theta)$ and $\mathcal{E}_n^G(\Theta)$. For P_f^n the product probability measure for density estimation and Q_f^n the law of the Gaussian white noise model* (1)*,*

$$\|P_{f_n}^n - P_{g_n}^n\|_{\mathrm{TV}} \to 0 \quad \Leftrightarrow \quad \|Q_{f_n}^n - Q_{g_n}^n\|_{\mathrm{TV}} \to 0 \quad \Leftrightarrow \quad n\int(\sqrt{f_n} - \sqrt{g_n})^2 \to 0 \tag{5}$$

and

$$\|P_{f_n}^n - P_{g_n}^n\|_{\mathrm{TV}} \to 1 \quad \Leftrightarrow \quad \|Q_{f_n}^n - Q_{g_n}^n\|_{\mathrm{TV}} \to 1 \quad \Leftrightarrow \quad n\int(\sqrt{f_n} - \sqrt{g_n})^2 \to \infty. \tag{6}$$

If $H^2(P,Q) = \int(\sqrt{dP} - \sqrt{dQ})^2$ denotes the Hellinger distance, then for $n > 1$,

$$H^2\big(Q_{f_n}^n, Q_{g_n}^n\big) \le H^2\big(P_{f_n}^n, P_{g_n}^n\big) \le H^2\big(Q_{f_n}^n, Q_{g_n}^n\big) + \frac{2\log n}{n}. \tag{7}$$

Proof. We first prove (7). By Lemma 5 below, $H^2(Q_{f_n}^n, Q_{g_n}^n) = 2 - 2\exp(-\frac{n}{2}\|\sqrt{f_n} - \sqrt{g_n}\|_2^2)$. Together with Lemmas 2.17 and 2.19 of [24], this proves

$$H^2\big(Q_{f_n}^n, Q_{g_n}^n\big) \le H^2\big(P_{f_n}^n, P_{g_n}^n\big) \le H^2\big(Q_{f_n}^n, Q_{g_n}^n\big) + \frac{1}{2}\int(\sqrt{f_n} - \sqrt{g_n})^2.$$

Distinguishing whether the term $\frac{n}{2}\int(\sqrt{f_n} - \sqrt{g_n})^2$ is larger or smaller than $\log n$, and using that $H^2(Q_{f_n}^n, Q_{g_n}^n) \ge 2 - 2n^{-1}$ if it is, then establishes (7).

To verify the first two assertions of the lemma, notice that by Le Cam's inequalities (Lemma 2.3 in [25]), for any probability measures P, Q,

$$\frac{1}{2}H^2(P,Q) \le \|P - Q\|_{\mathrm{TV}} \le \min\left\{H(P,Q), \left(1 - \frac{1}{2}\big(1 - \tfrac{1}{2}H^2(P,Q)\big)^2\right)\right\}. \tag{8}$$

Consequently, the total variation of two sequences (P_n) and (Q_n) converges to zero if and only if $H^2(P_n, Q_n) \to 0$. Similarly, $\|P_n - Q_n\|_{\mathrm{TV}} \to 1$ if and only if $H^2(P_n, Q_n) \to 2$. Using (7) and $H^2(Q_{f_n}^n, Q_{g_n}^n) = 2 - 2\exp(-\frac{n}{2}\|\sqrt{f_n} - \sqrt{g_n}\|_2^2)$, (5) and (6) follow. □

In view of this, we must construct sequences such that the total variation distances in the two experiments tend neither to zero nor one and are separated for n large enough.

In the following we describe the ideas that finally lead to a lower bound. As a first step, we use Lemma 4 below to show that in the density estimation model,

$$\|P_f^n - P_g^n\|_{\mathrm{TV}} \le 1 - \Big(1 - \frac{\|f-g\|_1}{2}\Big)^n.$$

In the Gaussian white noise model, we have by Lemma 5 that $\|Q_f^n - Q_g^n\|_{\mathrm{TV}} = 1 - 2\Phi(-\sqrt{n}\|\sqrt{f} - \sqrt{g}\|_2)$ with Φ the distribution function of a standard normal random variable. For the Le Cam deficiency of the binary experiments with parameter space $\Theta = \{f, g\}$, Lemma 1 then implies the following lower bound:

$$\delta\big(\mathcal{E}_n^D(\{f,g\}), \mathcal{E}_n^G(\{f,g\})\big) \ge \frac{1}{2}\Big(1 - \frac{\|f-g\|_1}{2}\Big)^n - \Phi\big(-\sqrt{n}\|\sqrt{f} - \sqrt{g}\|_2\big). \tag{9}$$

To prove asymptotic nonequivalence, we therefore want to construct sequences $(f_n)_n, (g_n)_n \subseteq \Theta$ such that the total variation $\|f_n - g_n\|_1$ is small while the Hellinger distance $\|\sqrt{f_n} - \sqrt{g_n}\|_2$ is large. The largest value of the Hellinger distance is given by Le Cam's inequalities (8), $\|\sqrt{f_n} - \sqrt{g_n}\|_2 \le \|f_n - g_n\|_1^{1/2}$. An inspection of the proof shows that equality is achieved if for all x, either $f_n(x) = 0$, $g_n(x) = 0$ or $f_n(x) = g_n(x)$. This is a first indication that the bound (9) is particularly useful for small densities.

We now provide a heuristic showing that the reduction to a binary experiment can only be used if the parameter space contains small densities. Observe that in view of Lemma 2, we need $\|f_n - g_n\|_1 \asymp \|\sqrt{f_n} - \sqrt{g_n}\|_2^2 \asymp 1/n$ to show that the Le Cam deficiency is lower bounded by a positive constant. The standard approach for nonparametric two hypothesis lower bounds is to consider one function as a local perturbation of the other. For fixed f_n and $K \not\equiv 0$ a smooth function with $\int K = 0$ and support in $[-1, 1]$, set

$$g_n = f_n + h_n^\beta K\Big(\frac{\cdot - x_0}{h_n}\Big), \tag{10}$$

where $x_0 \in (0,1)$ is fixed and $h_n > 0$. If f_n is a β-Hölder smooth function, a standard argument shows that g_n is also β-Hölder smooth. If the perturbation is small enough, then $g_n \ge 0$ is a density since $\int K = 0$. The perturbation has height $O(h_n^\beta)$ and support of length $2h_n$, which means that the total variation distance is of the order $h_n^{\beta+1}$. To ensure that $\|f_n - g_n\|_1 \asymp 1/n$, we therefore take $h_n \asymp n^{-1/(\beta+1)}$. On the other hand, the squared Hellinger distance satisfies

$$\|\sqrt{f_n} - \sqrt{g_n}\|_2^2 = \int \frac{(f_n - g_n)^2}{(\sqrt{f_n} + \sqrt{g_n})^2} \asymp \frac{h_n^{2\beta+1}}{f_n(x_0) + h_n^\beta}. \tag{11}$$

To ensure the right hand side is of order $1/n$, we consequently need $f(x_0) = O(h_n^\beta) = O(n^{-\frac{\beta}{\beta+1}})$. If all densities are bounded away from zero, a different approach based on a multiple testing problem is needed to obtain sharp lower bounds [22].

To summarize, we have used that in the density estimation model the total variation of the product measures P_f^n and P_g^n can be bounded in terms of the total variation between the densities f and g. On the contrary, in the Gaussian white noise model the total variation distance is a function of the Hellinger distance of f and g. The total variation distance is bounded from below by the *squared* Hellinger distance and from above by the Hellinger distance via (8). Nonequivalence can therefore be established using the inequality (9) if the total variation between f and g behaves like the squared Hellinger distance, which happens exactly when the densities are small.

3 Proofs

We construct two test functions and use that the Le Cam deficiency is bounded from below by the difference of the total variation distances. To prove Theorem 1, it is by (9) enough to show that for some densities $f_{1,n}, f_{2,n} \in \Theta_n$,

$$\begin{aligned}\frac{1}{2}\Big(1 - \frac{\|f_{1,n} - f_{2,n}\|_1}{2}\Big)^n - \Phi\big(-\sqrt{n}\|\sqrt{f_{1,n}} - \sqrt{f_{2,n}}\|_2\big) &\geq \frac{1}{2}e^{-\frac{3}{2}}\Big(1 - \sqrt{\frac{e}{\pi}}\Big) + o(1) \\ &\geq 0.007 + o(1). \qquad (12)\end{aligned}$$

We henceforth omit the index n for convenience, writing $f_1 = f_{1,n}$ and $f_2 = f_{2,n}$. Before we describe the construction of f_1, f_2, we first recall the following basic property of functions in the flat Hölder space $\mathcal{H}^\beta$.

Lemma 3 (Lemma 1 in [21]). *Suppose that $f \in \mathcal{H}^\beta$ with $\beta > 0$ and let $a = a(\beta) > 0$ be any constant satisfying $(e^a - 1) + a^\beta/(\lfloor\beta\rfloor!) \leq 1/2$. Then for*

$$|h| \leq a\left(\frac{|f(x)|}{\|f\|_{\mathcal{H}^\beta}}\right)^{1/\beta},$$

we have

$$|f(x+h) - f(x)| \leq \frac{1}{2}|f(x)|,$$

implying in particular, $|f(x)|/2 \leq |f(x+h)| \leq 3|f(x)|/2$.

Construction of $f_1, f_2 \in \Theta_n$: For a given density f_0, we consider two perturbations of f_0 for an x_0 such that $f_0(x_0) \asymp n^{-\beta/(\beta+1)}$. This choice is natural in view of (11). The way we construct the perturbations is for technical convenience slightly different than in (10). By

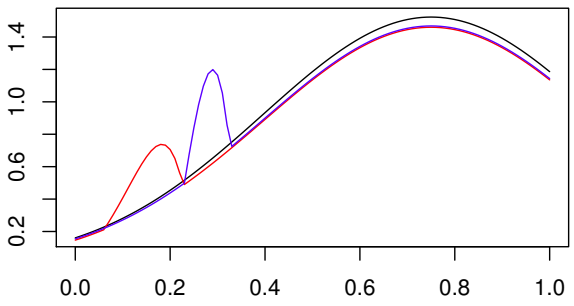

Figure 1: Plot of the densities f_0 (black), f_1 (red), f_2 (blue).

assumption there exist densities $f_0 := f_{0,n} \in \mathcal{H}^\beta(R)$ such that for some $x_0 := x_{0,n} \in [0,1]$, $R^{1/(\beta+1)}(2n)^{-\frac{\beta}{\beta+1}} \leq f_0(x_0) \leq R^{1/(\beta+1)} n^{-\frac{\beta}{\beta+1}}$ for all n. Without loss of generality, we may assume that $x_0 \leq 1/2$. We must ensure that we can apply Lemma 3 on the support of the perturbations, which motivates the following definitions. With $0 < a \leq 1/4$ a solution of $e^a - 1 + a^\beta/\lfloor\beta\rfloor! \leq 1/2$, set

$$F := \frac{a f_0(x_0)^{\frac{\beta+1}{\beta}}}{4R^{\frac{1}{\beta}}} \leq \frac{1}{16n}$$

and observe that $F \geq a/(8n)$. Given x_0, pick $x_1 < x_2$ such that $\int_{x_0}^{x_1} f_0(x)dx = \int_{x_1}^{x_2} f_0(x)dx = F$. By Lemma 3, $\int_{x_0}^{x_0 + a f_0(x_0)^{1/\beta}/R^{1/\beta}} f_0(x)dx \geq 2F$, which implies

$$x_2 \leq x_0 + a f_0(x_0)^{1/\beta}/R^{1/\beta} \leq 1/2 + R^{-\frac{1}{\beta(\beta+1)}} n^{-\frac{\beta}{\beta+1}}, \tag{13}$$

so that $x_2 \leq 1$ for $n \geq n_0(R,\beta)$ large enough.

Let $K \in \mathcal{C}^\beta(\mathbb{R})$ be a non-negative function supported on $[0,1]$ and satisfying $\int_0^1 K(u)du = 1$. For γ the solution of $\sqrt{1+\gamma} := 1 + 1/\sqrt{nF}$, consider the two test functions

$$f_j(x) = f_0(x)\left(1 - \gamma F + \gamma K\left(\frac{F_0(x) - F_0(x_{j-1})}{F}\right)\right), \quad j \in \{1,2\}, \tag{14}$$

where F_0 is the distribution function of f_0. Figure 1 displays an example of this construction. Since $\gamma = 1/(nF) + 2/\sqrt{nF}$, it follows that $\gamma F \leq 3/(2n)$ and $1 - \gamma F > 0$ for $n \geq 2$. By substitution, $\int_0^1 f_j(x)dx = 1$ and thus the f_j are densities. Moreover, $f_1 - f_0$ and $f_2 - f_0$ have disjoint support. We also have the following proposition that is proved below.

Proposition 2. *There exists a finite constant C, not depending on R, such that $f_1, f_2 \in \mathcal{H}^\beta(CR)$.*

Using $\gamma F \le 3/(2n)$ and $\gamma \le (1+\sqrt{8/a})^2$ gives

$$1 - \frac{3}{2n} \le \frac{f_j(x)}{f_0(x)} \le 1 + (1+\sqrt{8/a})^2 \|K\|_\infty.$$

It therefore follows that $f_1, f_2 \in \{f \in \mathcal{H}^\beta(cR) : c^{-1} f_0 \le f \le c f_0\}$ for

$$c = \max\big(C, 4, 1 + (1+\sqrt{8/a})^2 \|K\|_\infty\big) \tag{15}$$

and thus by assumption $f_1, f_2 \in \Theta_n$. We now establish (12) for these $f_1, f_2 \in \Theta_n$.

Lower bound for $\frac{1}{2}(1 - \frac{1}{2}\|f_1 - f_2\|_1)^n$: Using $\int_0^1 K(u)du = 1$ and substituting $u = (F_0(x) - F_0(x_{j-1}))/F$, we get $\|f_1 - f_2\|_1 = 2\gamma F$. Since $\gamma F \le 3/(2n)$,

$$\frac{1}{2}\Big(1 - \frac{\|f_1 - f_2\|_1}{2}\Big)^n \ge \frac{1}{2}\Big(1 - \frac{3}{2n}\Big)^n \to \frac{1}{2} e^{-\frac{3}{2}}. \tag{16}$$

Upper bound for $\Phi(-\sqrt{n}\|\sqrt{f_1} - \sqrt{f_2}\|_2)$: This is equivalent to lower bounding $\|\sqrt{f_1} - \sqrt{f_2}\|_2$. Splitting the integral $\int_0^1$ into $\int_{x_0}^{x_1} + \int_{x_1}^{x_2} + \int_{[x_0,x_2]^c}$, using the properties of K, substitution and the Cauchy-Schwarz inequality yields

$$\begin{aligned}
\|\sqrt{f_1} - \sqrt{f_2}\|_2^2 &= 2F \int_0^1 \big(\sqrt{1 - \gamma F + \gamma K(u)} - \sqrt{1 - \gamma F}\big)^2 du \\
&\ge 2F(1 - \gamma F) \int_0^1 \big(\sqrt{1 + \gamma K(u)} - 1\big)^2 du \\
&= 2F(1 - \gamma F)\big(\gamma + 2 - 2\int_0^1 \sqrt{1 + \gamma K(u)} du\big) \\
&\ge 2F(1 - \gamma F)\big(\sqrt{1+\gamma} - 1\big)^2 \\
&= \frac{2 - 2\gamma F}{n} \\
&\ge \frac{2}{n} - \frac{3}{n^2}
\end{aligned}$$

where in the last two lines we have used the definition of γ and $\gamma F \le 3/(2n)$. For $x > 0$, we find using the standard Gaussian tail bound $\Phi(-x) = 1 - \Phi(x) \le (2\pi)^{-1/2} x^{-1} e^{-x^2/2}$, so that we finally obtain

$$\Phi\big(-\sqrt{n}\big\|\sqrt{f_1} - \sqrt{f_2}\big\|_2\big) \le \Phi\big(-\sqrt{2}(1 + o(1))\big) \to \Phi\big(-\sqrt{2}\big) \le \frac{1}{2e\sqrt{\pi}}.$$

The last bound and (16) together imply (12), which completes the proof of Theorem 1. □

Proof of Proposition 2. In this proof we write $a_n \lesssim b_n$ if $a_n \le C b_n$ for a constant C which does not depend on R, but might depend on β and K. Note that $\gamma \le (1+\sqrt{8/a})^2$. Recall

that the support of $f_j - f_0$ is $[x_{j-1}, x_j]$ and that by Lemma 3, $\frac{1}{2}f_0(x) \le f_0(x_0) \le 2f_0(x)$ for all $x \in [x_0, x_2]$. The sup-norm can be easily bounded by $\|f_j\|_\infty \le \|f_0\|_\infty(1+\gamma\|K\|_\infty) \lesssim R$.

For $0 < \beta \le 1$, using the definition of F,

$$\begin{aligned}
|f_j|_{\mathcal{C}^\beta} &\le |f_0|_{\mathcal{C}^\beta}\big(1+\gamma\|K\|_\infty\big) + 2\gamma f_0(x_0)\left|K\left(\frac{F_0(\cdot)-F_0(x_{j-1})}{F}\right)\right|_{\mathcal{C}^\beta} \\
&\le R\big(1+\gamma\|K\|_\infty\big) + 2\gamma f_0(x_0)|K|_{\mathcal{C}^\beta}\sup_{x,y\in[x_{j-1},x_j]:x\neq y}\frac{|F_0(x)-F_0(y)|^\beta}{F^\beta|x-y|^\beta} \\
&\le R(1+\gamma\|K\|_\infty) + 2^{\beta+1}\gamma f_0(x_0)^{\beta+1}|K|_{\mathcal{C}^\beta}F^{-\beta} \\
&\le R(1+\gamma\|K\|_\infty + 16a^{-\beta}\gamma|K|_{\mathcal{C}^\beta}).
\end{aligned}$$

Since $|f_j|_{\mathcal{H}^\beta} = 0$ for $0 < \beta \le 1$ by definition, it follows that $\|f_j\|_{\mathcal{H}^\beta} \lesssim R$.

We now bound $|f_j|_{\mathcal{C}^\beta}$ for $\beta > 1$. Since $|f_0(1-\gamma F)|_{\mathcal{C}^\beta} \le R$, it remains to show $|f_0\cdot(K\circ v_j)|_{\mathcal{C}^\beta} \lesssim R$ with $v_j(x) := (F_0(x) - F_0(x_{j-1}))/F$. Let $1 \le r \le \lfloor\beta\rfloor$. For two r-times differentiable functions g, h, $(gh)^{(r)} = \sum_{q=0}^r \binom{r}{q} g^{(q)}h^{(r-q)}$. Moreover, by Faà di Bruno's formula,

$$\begin{aligned}
\big(K\circ v_j\big)^{(q)} &= \sum \frac{q!}{m_1!\dots m_q!}(K^{(M_q)}\circ v_j)\prod_{s=1}^q\left(\frac{v_j^{(s)}}{s!}\right)^{m_s} \\
&= \sum c_{m_1,\dots,m_q}\frac{K^{(M_q)}\circ v_j}{F^{M_q}}\prod_{s=1}^q\big(f_0^{(s-1)}\big)^{m_s},
\end{aligned}$$

where the sum is over all non-negative integers $m_1, \dots, m_q$ with $m_1 + 2m_2 + \dots + qm_q = q$ and $M_q := \sum_{s=1}^q m_s$. The r-th derivative of $f_0 \cdot (K \circ v_j)$ therefore equals

$$(K\circ v_j)f_0^{(r)} + \sum_{q=1}^r\binom{r}{q}\sum c_{m_1,\dots,m_q}\frac{K^{(M_q)}\circ v_j}{F^{M_q}}f_0^{(r-q)}\prod_{s=1}^q\big(f_0^{(s-1)}\big)^{m_s}, \tag{17}$$

where the second sum is over the same set of integers as above. We bound the $|\cdot|_{\mathcal{C}^\beta}$-seminorm by proving a Hölder bound for each of the terms in (17) individually, starting with the terms in the sum. For $1 \le q \le r$ and $\mathbf{m}_q = (m_1, \dots, m_q)$ a q-tuple as above, write

$$\varphi(x) = \varphi_{\mathbf{m}_q}(x) = F^{-M_q}(K^{(M_q)}\circ v_j)(x)f_0^{(r-q)}(x)\prod_{s=1}^q f_0^{(s-1)}(x)^{m_s},$$

so that we wish to prove $|\varphi(x) - \varphi(y)| \lesssim R|x-y|^{\beta-\lfloor\beta\rfloor}$. For any $x, y \in [x_0, x_2]$,

$$\begin{aligned}
|\varphi(x)-\varphi(y)| \le{}& \frac{1}{F^{M_q}}\left|K^{(M_q)}\circ v_j(x) - K^{(M_q)}\circ v_j(y)\right|\left|f_0^{(r-q)}(x)\prod_{s=1}^q f_0^{(s-1)}(x)^{m_s}\right| \\
&+ \frac{1}{F^{M_q}}\left|K^{(M_q)}\circ v_j(y)\left(f_0^{(r-q)}(x)\prod_{s=1}^q f_0^{(s-1)}(x)^{m_s} - f_0^{(r-q)}(y)\prod_{s=1}^q f_0^{(s-1)}(y)^{m_s}\right)\right|.
\end{aligned} \tag{18}$$

Using the definition (3) of $\mathcal{H}^\beta$ and that $\sum_{s=1}^q sm_s = q$,

$$\begin{aligned}\left|f_0^{(r-q)}(x)\prod_{s=1}^q f_0^{(s-1)}(x)^{m_s}\right| &\leq R^{\frac{r-q}{\beta}} f_0(x)^{\frac{\beta-(r-q)}{\beta}} \prod_{s=1}^q R^{\frac{(s-1)m_s}{\beta}} f_0(x)^{\frac{(\beta-s+1)m_s}{\beta}} \\ &= R^{\frac{r-M_q}{\beta}} f_0(x)^{\frac{\beta-r+(\beta+1)M_q}{\beta}}.\end{aligned} \tag{19}$$

Observe that by (13) and the definition of F, for any $1 \leq \ell < \beta$ and $x, y \in [x_0, x_2]$,

$$\begin{aligned}|K^{(\ell)}(v_j(x)) - K^{(\ell)}(v_j(y))| &\lesssim f_0(x_0)F^{-1}|x-y|^{\beta-\lfloor\beta\rfloor}|x_2-x_0|^{1-(\beta-\lfloor\beta\rfloor)} \\ &\lesssim R^{\frac{\beta-\lfloor\beta\rfloor}{\beta}} f_0(x_0)^{\frac{\lfloor\beta\rfloor-\beta}{\beta}}|x-y|^{\beta-\lfloor\beta\rfloor}.\end{aligned}$$

Combining the previous two displays and again using the definition of F, the first term in (18) is $O(R^{\frac{\beta-\lfloor\beta\rfloor+r}{\beta}} f_0(x_0)^{\frac{\lfloor\beta\rfloor-r}{\beta}}|x-y|^{\beta-\lfloor\beta\rfloor}) = O(R|x-y|^{\beta-\lfloor\beta\rfloor})$ as required.

For $1 \leq \ell < \lfloor\beta\rfloor$, $m \in \mathbb{N}$ and $x, y \in [x_0, x_2]$ with $x < y$, we have using (3), (13) and the mean value theorem, that for some $\xi = \xi_{x,y} \in [x, y]$,

$$\begin{aligned}|f_0^{(\ell)}(x)^m - f_0^{(\ell)}(y)^m| &\leq m|f_0^{(\ell+1)}(\xi)||f_0^{(\ell)}(\xi)|^{m-1}(x_2-x_0)^{1-(\beta-\lfloor\beta\rfloor)}|x-y|^{\beta-\lfloor\beta\rfloor} \\ &\lesssim R^{\frac{\beta-\lfloor\beta\rfloor+\ell m}{\beta}} f_0(x_0)^{\frac{\lfloor\beta\rfloor-\beta+(\beta-\ell)m}{\beta}}|x-y|^{\beta-\lfloor\beta\rfloor}.\end{aligned}$$

Noting that for $\ell = \lfloor\beta\rfloor$, m_ℓ only takes values 0 or 1 in the sum in (17), one can extend the previous display to $\ell = \lfloor\beta\rfloor$ by directly using the Hölder continuity of $f_0^{(\lfloor\beta\rfloor)}$. For the second term in (18), we repeatedly apply the triangle inequality, each time changing the variable in a single derivative. Fix an integer $1 \leq k \leq q$ and define vectors $(z_s)_{s=1}^q$ and $(\tilde{z}_s)_{s=1}^q$, which are identically equal to x or y in all entries except the k^{th}-coordinate, where $z_k = x$ and $\tilde{z}_k = y$. Then using the previous display, a similar argument to (19) and the definition of F,

$$\begin{aligned}&\frac{1}{F^{M_q}}\left|f_0^{(r-q)}(x)\right|\left|\prod_{s=1}^q f_0^{(s-1)}(z_s)^{m_s} - \prod_{s=1}^q f_0^{(s-1)}(\tilde{z}_s)^{m_s}\right| \\ &= \frac{1}{F^{M_q}}\left|f_0^{(r-q)}(x)\right|\left|\prod_{s=1, s\neq k}^q f_0^{(s-1)}(z_s)^{m_s}\right|\left|f_0^{(k-1)}(x)^{m_k} - f_0^{(k-1)}(y)^{m_k}\right| \\ &\lesssim R^{\frac{\beta-\lfloor\beta\rfloor+r}{\beta}} f_0(x_0)^{\frac{\lfloor\beta\rfloor-r}{\beta}}|x-y|^{\beta-\lfloor\beta\rfloor} \\ &\lesssim R|x-y|^{\beta-\lfloor\beta\rfloor},\end{aligned}$$

where is the last line we have used that $r \leq \lfloor\beta\rfloor$ and $\|f_0\|_\infty \leq R$. The same inequality can be established for $F^{-M_q}|f_0^{(r-q)}(x) - f_0^{(r-q)}(y)||\prod_{s=1}^q f_0^{(s-1)}(y)^{m_s}|$. Since $\|K^{(M_q)} \circ v_j\|_\infty \lesssim 1$, by repeatedly applying the triangle inequality and the last display, the second term in (18) is $O(R|x-y|^{\beta-\lfloor\beta\rfloor})$, so that $|\varphi(x)-\varphi(y)| \lesssim R|x-y|^{\beta-\lfloor\beta\rfloor}$ as required. A similar, but simpler,

argument shows that the first term in (17) satisfies $|K \circ v_j(x) f_0^{(r)}(x) - K \circ v_j(y) f_0^{(r)}(y)| \lesssim R|x-y|^{\beta - \lfloor \beta \rfloor}$. This shows that every term in (17), and hence the whole of (17), satisfies the required Hölder bound, so that $|f_0 \cdot (K \circ v_j)|_{\mathcal{C}^\beta} \lesssim R$.

We now prove that $|f_j|_{\mathcal{H}^\beta} \lesssim R$ for $\beta > 1$. By (3), it suffices to show $|f_j^{(r)}(x)| \leq (CR)^{\frac{r}{\beta}} |f_j(x)|^{\frac{\beta - r}{\beta}}$ for all $x \in [0,1]$ and $r = 1, \ldots, \lfloor \beta \rfloor$ and a constant C that does not depend on R. Since $K \geq 0$ and $\gamma F \leq 3/(2n)$, it is enough to show that $|f_j^{(r)}(x)| \leq (C'R)^{\frac{r}{\beta}} |f_0(x)|^{\frac{\beta - r}{\beta}}$ for all $x \in [0,1]$ and a possibly different constant C'. This follows if $|(f_0 \gamma (K \circ v_j))^{(r)}(x)| \leq (C''R)^{\frac{r}{\beta}} |f_0(x)|^{\frac{\beta - r}{\beta}}$ for all $x \in [0,1]$, $r = 1, \ldots, \lfloor \beta \rfloor$ and some $C'' < \infty$. This last inequality follows from (17), (19), the definition of F and that $f_0 \in \mathcal{H}^\beta(R)$. This also shows that $\|f_j^{(\lfloor \beta \rfloor)}\|_\infty \lesssim R^{\frac{\lfloor \beta \rfloor}{\beta}} \|f_0\|_\infty^{\frac{\beta - \lfloor \beta \rfloor}{\beta}} \lesssim R$. □

Lemma 4. *For P and Q dominated probability measures, the product measures $P^{\otimes n} = P \otimes \cdots \otimes P$ and $Q^{\otimes n} = Q \otimes \cdots \otimes Q$ satisfy*

$$\left\| P^{\otimes n} - Q^{\otimes n} \right\|_{\mathrm{TV}} \leq 1 - \left(1 - \|P - Q\|_{\mathrm{TV}}\right)^n.$$

Proof. For probability measures $\widetilde{P}, \widetilde{Q}$ on the same measurable space, we have $\|\widetilde{P} - \widetilde{Q}\|_{\mathrm{TV}} = 1 - \int d\widetilde{P} \wedge \widetilde{Q}$. If p, q denote the densities of P, Q with respect to some dominating measure ν,

$$\left\| P^{\otimes n} - Q^{\otimes n} \right\|_{\mathrm{TV}} = 1 - \int \prod_{i=1}^n p(x_i) \wedge \prod_{i=1}^n q(x_i) d\nu(x_i) \leq 1 - \prod_{i=1}^n \int p(x_i) \wedge q(x_i) d\nu(x_i)$$

and the right hand side can be rewritten as $1 - (1 - \|P - Q\|_{\mathrm{TV}})^n$. □

Lemma 5. *For a function b and $\sigma > 0$, denote by $Q_{b,\sigma}$ the distribution of the path $(Y_t)_{t \in [0,1]}$ with $dY_t = b(t)dt + \sigma dW_t$, where W is a Brownian motion. If Φ denotes the distribution function of the standard normal distribution, then*

$$\|Q_{b_1,\sigma} - Q_{b_2,\sigma}\|_{\mathrm{TV}} = 1 - 2\Phi(-\tfrac{1}{2\sigma} \|b_1 - b_2\|_2),$$

$$H^2(Q_{b_1,\sigma}, Q_{b_2,\sigma}) = 2 - 2\exp(-\tfrac{1}{8\sigma^2} \|b_1 - b_2\|_2^2).$$

Proof. This follows from Girsanov's formula $dQ_{b,\sigma}/dQ_{0,\sigma} = \exp(\sigma^{-1} \int_0^1 b(t) dW_t - \frac{1}{2}\sigma^{-2} \|b\|_2^2)$ together with $\|P - Q\|_{\mathrm{TV}} = 1 - P(\frac{dQ}{dP} > 1) - Q(\frac{dP}{dQ} \geq 1)$ and $H^2(P,Q) = 2 - 2\int (dP dQ)^{1/2}$. □

Proof of Proposition 1. We verify the conditions of Theorem 1.2 of Nussbaum [18], which is a specialized version of the more general Theorem 4.3 of Le Cam [13]. We first begin with

the regularity conditions stated in Section 10 of [18]. Since $g'(x) = 960x(1/2-x)(1-4x)$, for any $\theta \in (0, 1/2)$ the Fisher information equals

$$I(\theta) = \int_\theta^{\theta+1/2} \frac{g'(x-\theta)^2}{g(x-\theta)} dx = \int_0^1 \frac{g'(x)^2}{g(x)} dx = 960 \int_0^{1/2} (1-4x)^2 dx = 160,$$

from which we observe that $I(\theta)$ is both constant and finite for all $\theta \in (0, 1/2)$. For $\dot{\ell}_\theta = f_\theta^{-1} \frac{\partial}{\partial \theta} f_\theta$, one can show explicitly that for $\theta, \theta + h \in (0, 1/2)$,

$$\int_0^1 \left[\sqrt{f_{\theta+h}} - \sqrt{f_\theta} - \tfrac{1}{2} h \dot{\ell}_\theta \sqrt{f_\theta} \right]^2 = 960 h^4,$$

so that the family $(f_\theta : \theta \in K)$ is differentiable in quadratic mean uniformly on compact sets $K \subset (0, 1/2)$ in the sense of p. 578 of Le Cam [14]. Together, these verify the regularity conditions of Proposition 1.2 of [18], see Section 10 of [18] or Proposition 1, Chapter 17.3 of [14].

We now verify the crucial condition that the family $\Theta(K)$ has finite Hellinger metric dimension. Using the explicit form of g and directly integrating, one can show after some calculations that for any $\theta \in (0, 1/2)$,

$$\int_0^1 \sqrt{g(x-\theta)g(x)} dx = \int_\theta^{1/2} \sqrt{g(x-\theta)g(x)} dx = 1 - 20\theta^2(1 - 2\theta + 8\theta^3/5),$$

so that the Hellinger distance equals

$$H^2(P_{f_\theta}, P_g) = 2 - 2 \int_0^1 \sqrt{g(x-\theta)g(x)} dx = 40\theta^2(1 - 2\theta + 8\theta^3/5).$$

By a change of variable, $H(P_{f_\theta}, P_{f_{\theta'}}) = H(P_{f_{|\theta-\theta'|}}, P_g)$ for any $\theta, \theta' \in (0, 1/2)$, so that $H^2(P_{f_\theta}, P_{f_{\theta'}}) = 40|\theta - \theta'|^2 + O(|\theta - \theta'|^3)$. In particular, there exists a constant $\tilde{c} > 1$ such that $\tilde{c}^{-1}|\theta - \theta'|^2 \leq H^2(P_{f_\theta}, P_{f_{\theta'}}) \leq \tilde{c}|\theta - \theta'|^2$ for all $\theta, \theta' \in (0, 1/2)$.

To establish finite Hellinger metric dimensionality, we must show that for any $\varepsilon > 0$, $\{f_{\theta'} : H(f_{\theta'}, f_\theta) \leq \varepsilon\}$ can be covered in Hellinger distance by a finite number of $\varepsilon/2$ balls, independently of ε. By the above results, it suffices to show that $\{\theta' : |\theta' - \theta| \leq \tilde{c}^{1/2}\varepsilon\}$ can be covered by N balls $\{\theta' : |\theta' - \theta_i| \leq \tilde{c}^{-1/2}\varepsilon/2\}$, $i = 1, ..., N$, for some N independent of ε. This is simply covering a compact interval in $\mathbb{R}$ and can be done with $N = 2\tilde{c}$ such $\varepsilon/2$ balls, thereby giving the required finite metric dimension $D \leq \log_2(2\tilde{c})$. □

Acknowledgements: The authors would like to thank the referees for their helpful comments and a referee from another paper for suggesting the example in Proposition 1. Most of this work was done while Kolyan Ray was a postdoctoral researcher at Leiden University.

References

[1] Brown, L., Cai, T., Zhang, R., Zhao, L., and Zhou, H. The root–unroot algorithm for density estimation as implemented via wavelet block thresholding. *Probab. Theory Related Fields 146*, 3 (2009), 401–433.

[2] Brown, L. D., Carter, A. V., Low, M. G., and Zhang, C.-H. Equivalence theory for density estimation, Poisson processes and Gaussian white noise with drift. *Ann. Statist. 32*, 5 (10 2004), 2074–2097.

[3] Brown, L. D., and Low, M. G. Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist. 24*, 6 (12 1996), 2384–2398.

[4] Brown, L. D., and Zhang, C.-H. Asymptotic nonequivalence of nonparametric experiments when the smoothness index is 1/2. *Ann. Statist. 26*, 1 (1998), 279–287.

[5] Dalalyan, A., and Reiss, M. Asymptotic statistical equivalence for scalar ergodic diffusions. *Probab. Theory Related Fields 134*, 2 (2006), 248–282.

[6] Dalalyan, A., and Reiss, M. Asymptotic statistical equivalence for ergodic diffusions: the multidimensional case. *Probab. Theory Related Fields 137*, 1-2 (2007), 25–47.

[7] Delattre, S., and Hoffmann, M. Asymptotic equivalence for a null recurrent diffusion. *Bernoulli 8*, 2 (2002), 139–174.

[8] Efromovich, S., and Samarov, A. Asymptotic equivalence of nonparametric regression and white noise model has its limits. *Statist. Probab. Lett. 28*, 2 (1996), 143–145.

[9] Genon-Catalot, V., Laredo, C., and Nussbaum, M. Asymptotic equivalence of estimating a Poisson intensity and a positive diffusion drift. *Ann. Statist. 30*, 3 (2002), 731–753.

[10] Golubev, G. K., Nussbaum, M., and Zhou, H. H. Asymptotic equivalence of spectral density estimation and Gaussian white noise. *Ann. Statist. 38*, 1 (2010), 181–214.

[11] Grama, I., and Nussbaum, M. Asymptotic equivalence for nonparametric generalized linear models. *Probab. Theory Related Fields 111*, 2 (1998), 167–214.

[12] Hohage, T., and Werner, F. Inverse problems with Poisson data: statistical regularization theory, applications and algorithms. *Inverse Problems 32*, 9 (2016), 093001, 56.

[13] Le Cam, L. Sur l'approximation de familles de mesures par des familles gaussiennes. *Ann. Inst. H. Poincaré Probab. Statist. 21*, 3 (1985), 225–287.

[14] Le Cam, L. *Asymptotic methods in statistical decision theory.* Springer Series in Statistics. Springer-Verlag, New York, 1986.

[15] Low, M. G., and Zhou, H. H. A complement to Le Cam's theorem. *Ann. Statist.*

35, 3 (2007), 1146–1165.

[16] Makitalo, M., and Foi, A. Optimal inversion of the Anscombe transformation in low-count Poisson image denoising. *IEEE Trans. Image Process. 20*, 1 (2011), 99–109.

[17] Mariucci, E. Asymptotic equivalence for density estimation and Gaussian white noise: an extension. *Ann. I.S.U.P. 60*, 1-2 (2016), 23–34.

[18] Nussbaum, M. Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist. 24*, 6 (12 1996), 2399–2430.

[19] Patschkowski, T., and Rohde, A. Adaptation to lowest density regions with application to support recovery. *Ann. Statist. 44*, 1 (2016), 255–287.

[20] Ray, K., and Schmidt-Hieber, J. Minimax theory for a class of nonlinear statistical inverse problems. *Inverse Problems 32*, 6 (2016), 065003.

[21] Ray, K., and Schmidt-Hieber, J. A regularity class for the roots of nonnegative functions. *Ann. Mat. Pura Appl. (4) 196*, 6 (2017), 2091–2103.

[22] Ray, K., and Schmidt-Hieber, J. The Le Cam distance between density estimation, Poisson processes and Gaussian white noise. *Mathematical Statistics and Learning* (2018). to appear.

[23] Schmidt-Hieber, J. Asymptotic equivalence for regression under fractional noise. *Ann. Statist. 42*, 6 (2014), 2557–2585.

[24] Strasser, H. *Mathematical theory of statistics*, vol. 7 of *De Gruyter Studies in Mathematics*. Walter de Gruyter & Co., Berlin, 1985.

[25] Tsybakov, A. B. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009.

[26] Wang, Y. Asymptotic nonequivalence of Garch models and diffusions. *Ann. Statist. 30*, 3 (2002), 754–783.